(4)

The Pennsylvania State University

Department of Statistics

University Park, Pennsylvania

TECHNICAL REPORTS AND PREPRINTS

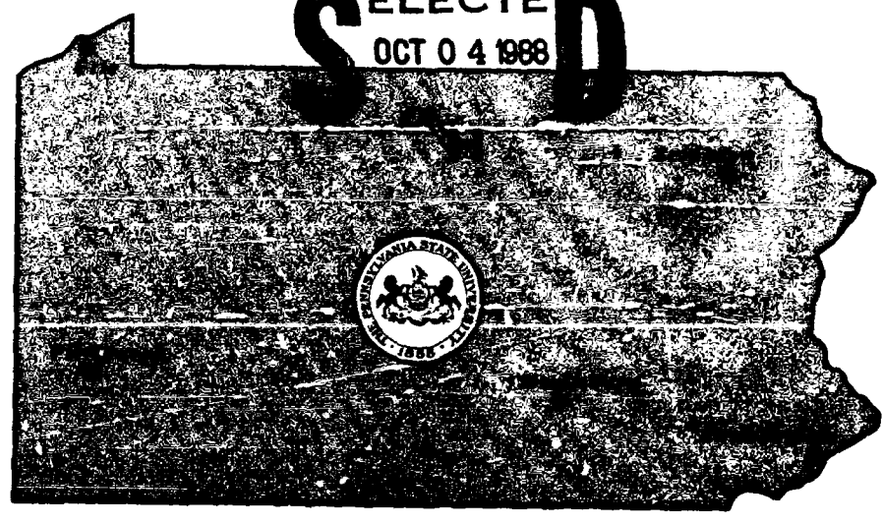
Number 76: September 1988

ROBUST DIAGNOSTICS FOR RANK-BASED INFERENCE

Joseph W. McKean
Western Michigan University

Simon J. Sheather
University of New South Wales

Thomas P. Hettmansperger*
Pennsylvania State University



DEPARTMENT OF STATISTICS

The Pennsylvania State University
University Park, PA 16802 U.S.A.

TECHNICAL REPORTS AND PREPRINTS

Number 76: September 1988

ROBUST DIAGNOSTICS FOR RANK-BASED INFERENCE

Joseph W. McKean
Western Michigan University

Simon J. Sheather
University of New South Wales

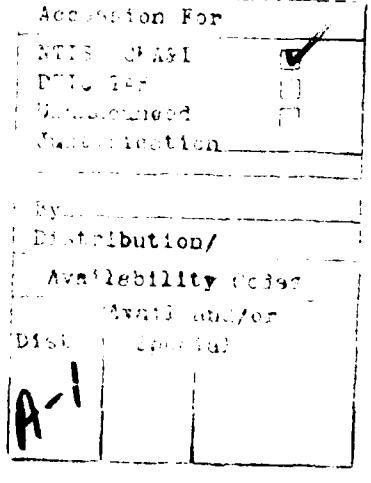
Thomas P. Hettmansperger*
Pennsylvania State University

*Research partially supported by ONR Contract N00014-80-C0741.

Abstract

Diagnostics based on robust R-estimates of regression coefficients are developed. These methods are not as sensitive to influential points as least squares diagnostics. In data sets with several influential points, diagnostics based on a robust fit have a greater chance of detecting interesting cases for further inspection. Robust analogues of the internal and external t statistics, DFFITS, DCOOK, and DFBETAS are developed and illustrated on two data sets.

Keys words: Linear Models, Robustness, Regression.



1. Introduction

A regression model is at best an approximation to the reality of the situation under study. Regression diagnostics are invaluable tools for detecting data points at which the model and the data differ greatly (such points are called outliers) as well as data points which have a large influence on the model. In the last ten years there has been much interest in the area of regression diagnostics. A testament to this is that a number of diagnostics are currently available in all the major statistical computing packages. Regression diagnostics are discussed in detail in the books by Cook and Weisberg (1982) and Belsley, Kuh and Welsch (1980) and in the review articles by Hocking (1983), Chatterjee and Hadi (1986), and Hettmansperger (1987).

It is well known, however, that a few influential points can spoil the least squares fit of a linear model. In data sets with several influential points, some of these points can exert such a strong influence on the least squares fit that other influential points are masked and, hence, are not detected by these diagnostic procedures. The data sets discussed in Section 4 are illustrations of this effect. Examination of data sets containing influential points, based on estimates that are impaired by such points, is a serious drawback to diagnostics based on least squares. In these circumstances, the traditional diagnostics suffer a lack of detection power.

Over the last ten years, the area of robust regression has also become a rapidly expanding field. Some of the major statistical packages now contain some form of robust regression. Using a robust fitting method reduces the effect of influential points on the fitted model. However, a number of authors point out that the exclusive use of robust methods can obscure important substantive problems with the model which in some situations are revealed by regression diagnostics based on least squares; see Cook (1986) and Chatterjee and Hadi (1986).

In this paper we develop diagnostics based on robust R-estimates of regression coefficients. Similar methods can be used to develop diagnostics based on other classes of robust estimates. These estimates are not as sensitive to influential points as least squares and the resulting diagnostics

appear to be more powerful than the least squares based methods. In data sets with several influential points, the diagnostics based on the robust fit therefore have a greater chance of detecting influential points than those based on least squares. In Section 3, we develop robust analogues of the internal and external t statistic, DFFITS, DCOOK, and DFBETAS. As Sections 2 and 3 demonstrate, their geometry is quite similar to their least squares counterparts. The last four techniques measure the impact of an individual case on the robust fit. In the examples of Section 4, these four techniques are able to detect some obvious outliers whereas the same techniques based on least squares are not. The robust diagnostics can thus be used to flag potential cases of trouble and should serve as quite useful tools in linear model fitting.

In Appendix A, we present a unified development of some of the more useful least squares diagnostics. The derivations are based on the mean shift outlier model; see Cook and Weisberg (1982, p.20).

2. Notation and R-Estimates

Consider the linear model

$$\underline{Y} = \alpha\underline{1} + X_C\underline{\beta} + \underline{e} \qquad (2.1)$$

where $\underline{1}$ denotes an $n\times 1$ vector of ones, X_C is an $n\times(p-1)$ centered design matrix having full column rank, α is an intercept parameter, $\underline{\beta}$ is $(p-1)\times 1$ vector of parameters, and $\underline{e}$ is an $n\times 1$ vector of random errors whose components are i.i.d. with distribution function F and density f. Letting $X = [\underline{1}: X_C]$ and $\underline{b} = (\alpha,\underline{\beta}')'$, we can write the model as

$$\underline{Y} = X\underline{b} + \underline{e}.$$

Discussions of R-estimates for this linear model can be found in Hettmansperger and McKean (1977). Briefly, consider Jaeckel's (1972) dispersion function which is given by

$$D(\underline{\beta}) = \Sigma a(R(Y_i - \underline{x}_{Ci}'\underline{\beta}))(Y_i - \underline{x}_{Ci}'\underline{\beta}) \qquad (2.2)$$

where $\underline{x}_{Ci}'$ is the ith row of X_C, $R(u_i)$ denotes the rank of u_i among $u_1,\ldots,u_n$

and $\{a(i)\}$ is a set of scores which are generated as

$$a(i) = \varphi(\frac{i}{n+1}) \tag{2.3}$$

where φ is a nondecreasing function defined on (0,1) such that $\int\varphi(u)du = 0$ and $\int\varphi^2(u)du = 1$. Examples of such score functions are the Wilcoxon, $\varphi(u) = \sqrt{12}(u-\frac{1}{2})$, and the sign scores $\varphi(u) = sgn(u-\frac{1}{2})$.

Jaeckel (1972) showed that D is a continuous, convex function of $\underline{\beta}$ and proposed estimating $\underline{\beta}$ by $\hat{\underline{\beta}}_R$ where $D(\hat{\underline{\beta}}_R) = \min_{\underline{\beta}} D(\underline{\beta})$. McKean and Hettmansperger (1976) proposed testing subvectors of $\underline{\beta}$ by using the reduction in $D(\underline{\beta})$ due to fitting the full and reduced models. Algorithms for obtaining $\underline{\beta}_R$ can be found in McKean and Hettmansperger (1978) and Osborne (1985). Version 6 of MINITAB contains commands which return $\hat{\underline{\beta}}_R$.

Under mild regularity conditions, found in Heiler and Willers (1979), $\hat{\underline{\beta}}_R$ satisfies

$$\hat{\underline{\beta}}_R = \underline{\beta} + \tau(X_c'X_c)^{-1}X_c'\underline{a}(R(\underline{e})) + o_p(1) \tag{2.4}$$

where $\underline{a}(R(\underline{e}))$ denotes the vector with components $a(R(e_i))$ and τ is a scale parameter defined by

$$\tau = \int\varphi(u)(-\frac{f'(F^{-1}(u))}{f(F^{-1}(u))})du. \tag{2.5}$$

Discussions of consistent estimates of τ based on the residuals $\underline{Y} - X_c\hat{\underline{\beta}}_R$ can be found in Koul et al. (1987) and Aubuchon and Hettmansperger (1988). Under these regularity conditions $X_c'\underline{a}(R(\underline{e}))$ is approximately $N_{p-1}(\underline{0},X_c'X_c)$; hence,

$$\hat{\underline{\beta}}_R \text{ is approximately } N_{p-1}(\underline{\beta},\tau^2(X_c'X_c)^{-1}). \tag{2.6}$$

Note if $X_c = [X_{1c}|X_{2c}]$ and X_{1c} and X_{2c} are orthogonal, i.e. $X_{1c}'X_{2c} = 0$, then $\hat{\underline{\beta}}_{1R}$ and $\hat{\underline{\beta}}_{2R}$ are asymptotically independent. While this does not imply, in finite samples, that the R-estimates of $\underline{\beta}_1$ are the same in the reduced and

full models, we have found in practice that the estimates do not differ by very much. For use in Section 4, we will term this approximate orthogonality.

Also, since ranks are invariant to constant shifts, the intercept parameter cannot be estimated from $D(\underline{\beta})$. If symmetry of the error distribution seems to be a tenable assumption, then scores satisfying

$$\varphi(1-u) = -\varphi(u) \tag{2.7}$$

are suitable; for instance, both the Wilcoxon and sign scores cited above satisfy this condition. Then the intercept can be estimated by using a one sample location R-estimate which corresponds to the chosen score function, see McKean and Hettmansperger (1978). Let $\hat{\alpha}_R$ denote this estimate and let $\hat{\underline{b}}_R = (\hat{\alpha}_R, \hat{\underline{\beta}}_R')'$. Under regularity conditions which include symmetry of the distribution of the errors, $\hat{\underline{b}}_R$ is approximately $N_p(\underline{b}, \tau^2(X'X)^{-1})$.

If symmetry of the error distribution is not tenable then, to avoid its assumption, we take the intercept to be the median of the distribution $F(y-\underline{x}'\underline{\beta})$ and estimate the intercept by

$$\hat{\alpha}^* = \text{med}\{Y_i - \underline{x}_{ic}'\hat{\underline{\beta}}_R\}.$$

Under the same regularity conditions as cited for (2.4), we have

$$\hat{\alpha}^* = \alpha + \tau^* \frac{1}{n} \underline{1}'\underline{a}^*(\underline{e}) + o_p(1) \tag{2.8}$$

where $\tau^* = (2f(0))^{-1}$ and $a^*(e_i) = \text{sgn}(e_i)$. Estimation of τ^* is discussed by McKean and Schrader (1984). It then follows that

$$\begin{pmatrix} \hat{\alpha}^* \\ \hat{\underline{\beta}}_R \end{pmatrix} \text{ is approximately } N_p\left(\begin{pmatrix} \alpha \\ \underline{\beta} \end{pmatrix}, \begin{bmatrix} \tau^{*2} & \underline{0}' \\ \underline{0} & \tau^2(X_c'X_c)^{-1} \end{bmatrix}\right).$$

3. R-Diagnostics

3.1. Internal R-studentized residuals.

Similar to the least squares residuals, the variance of the R-residuals, $\hat{e}_{R,i}$, depend on both the linear model and the underlying variation of the errors. The internal studentized least squares residuals, see Appendix A, have proved useful in diagnostic procedures since they correct for both the model and the underlying variance. The internal R-studentized residuals defined below, (3.5), are similarly standardized R-residuals.

As discussed in Section 2, let $\hat{\alpha}^*$ and $\hat{\underline{\beta}}_R$ denote the R-estimates of α and $\underline{\beta}$. Denote the residuals by $\hat{\underline{e}}_R = \underline{Y} - \hat{\alpha}^*\underline{1} - X_C\hat{\underline{\beta}}_R$. In order to standardize these residuals we need an estimate of the variance-covariance matrix $\text{Cov}(\hat{\underline{e}}_R)$. From Appendix B, equation (B.9), we have the approximation

$$\hat{\underline{e}}_R \doteq (\underline{Y}-\underline{1}\alpha-X_C\underline{\beta}) - \tau^*J\underline{a}^* - \tau H_C\underline{a} \tag{3.1}$$

where $\underline{a}^*$ and $\underline{a}$ denote the vectors $\underline{a}^*(\underline{e})$ and $\underline{a}^*(R(\underline{e}))$ given in the expressions (2.4) and (2.8). Throughout this paper $\doteq$ refers to first order approximations as developed in Appendix B. As shown in (B.1) of Appendix B, an estimate of the variance-covariance matrix of $\hat{\underline{e}}_R$ is

$$\tilde{S} = \hat{\sigma}^2\{I-\hat{K}_1J-\hat{K}_2H_C\} \tag{3.2}$$

where $\hat{K}_1 = (\hat{\tau}^*/\hat{\sigma})^2((2\hat{\delta}^*/\hat{\tau}^*)-1)$

$\hat{K}_2 = (\hat{\tau}/\hat{\sigma})^2((2\hat{\delta}/\hat{\tau})-1)$

$\hat{\delta}^* = \frac{1}{n-p}\Sigma|\hat{e}_{R,i}|$

and $\hat{\delta} = \frac{1}{n-p}D(\hat{\underline{\beta}}_R)$.

The estimators $\hat{\tau}^*$ and $\hat{\tau}$ are discussed in Section 2 and $D(\hat{\beta}_R)$ is defined by (2.2).

To complete the estimate of the $Cov(\hat{\underline{e}}_R)$ we need an estimate of σ^2. One possibility is to use the least squares estimate $\hat{\sigma}^2$. This is a consistent estimate provided the errors have finite variance. There are other possibilities but they involve assumptions on the form of the distribution; for example, $\hat{\tau}\hat{\delta}$ is a consistent estimate provided the errors have a normal distribution. For robustness, a mildly trimmed or winsorized mean square error could be used, see Shoemaker and Hettmansperger (1982).

It follows from (3.2) that an estimate of $Var(\hat{e}_{R,i})$ is

$$\tilde{s}^2_{R,i} = \hat{\sigma}^2(1 - \hat{K}_1 \frac{1}{n} - \hat{K}_2 h_{ic}), \tag{3.3}$$

where $h_{ic} = \underline{x}'_{ic}(X'_cX_c)^{-1}\underline{x}_{ic}$.

Note that in the least squares case $\tilde{s}^2_{LS,i} = \hat{\sigma}^2(1-h_i)$ and $h_i = n^{-1} + h_{ic}$ the ith diagonal element of the least squares projection matrix, which is the ith leverage value. Hence $\hat{K}_1$ and $\hat{K}_2$ can be viewed as corrections due to using the rank based fitting method. If the error distribution is symmetric (3.3) reduces to

$$\tilde{s}^2_{R,i} = \hat{\sigma}^2(1-\hat{K}_2 h_i). \tag{3.4}$$

We define the internal R-studentized residuals as

$$r_{R,i} = \frac{\hat{e}_{R,i}}{\tilde{s}_{R,i}} \qquad i=1,\ldots,n \tag{3.5}$$

where $\tilde{s}_{R,i}$ is the square root of either (3.3) or (3.4) depending on whether one assumes an asymmetric or symmetric error distribution, respectively.

As with their least squares counterparts, we think the chief benefit of the internal R-studentized residuals is their usefulness in diagnostic plots, such as plots of residuals versus fitted values and q - q plots. These residuals are corrected for both the design and the underlying variance.

It is interesting to compare expression (3.4) with the estimate of the variance of the least squares residual, $\hat{\sigma}^2(1-h_i)$. The correction factor $\hat{K}_2$

depends on the score function $\varphi(\cdot)$ and the underlying symmetric error distribution. If, for example, the error distribution is normal and if we use normal scores, then $\hat{K}_2$ converges almost surely to 1. In general, however, we will not wish to specify the error distribution and then $\hat{K}_2$ provides a natural adjustment.

3.2. R-estimates for the mean shift outliers model.

The R diagnostics that follow depend on the mean shift outlier model which is discussed in detail in Appendix A. Briefly, for the ith case, the mean shift outlier model is

$$\underline{Y} = X\underline{b} + \underline{d}_i\theta_i + \underline{e} \tag{3.6}$$

where $\underline{d}_i$ is an n×1 vector of zeroes except for its ith component which is 1. A formal test that the ith point is an outlier involves testing the hypotheses H_0: $\theta_i=0$ versus H_A: $\theta_i\neq 0$.

Below we obtain an R-estimate of θ_i and an estimate $\hat{\tau}(i)$ of τ, based on the model (3.6). These estimates will play a key role for the R-diagnostics that follow.

One way of obtaining an R-estimate of θ_i involves fitting this model. This would be computationally expensive since n such models need to be fit. Another way would be to consider aligned rank procedures. These procedures remove the effects of nuisance parameters (in this case $\underline{b}$) by considering the residuals from the reduced model (in this case $\hat{\underline{e}}_R$ from the reduced model $\underline{Y} = X\underline{b} + \underline{e}$); see Puri and Sen (1985) for a discussion of aligned rank procedures.

It is convenient to use the second form of the mean shift outlier model (A.3) given by $\underline{Y} = X\underline{b}^* + \underline{d}_i^*\theta_i + \underline{e}$, where $\underline{d}_i^* = (I-H)\underline{d}_i$, and H is $X(X'X)^{-1}X'$. In this form X and $\underline{d}_i^*$ are orthogonal and McKean (1975) has shown that this helps eliminate bias in the estimates. This is the model Cook and Weisberg (1982) used in obtaining the least squares external t diagnostic. Note that

the first part of the model $X\underline{b}^*$ is a vector in the column space of X. Hence the R-residuals from the fit of this reduced model are still $\underline{e}_R$.

Our R-estimate of θ_i is $\hat{\theta}_{R,i}$ which is a solution of $A_i^*(\theta_i) = 0$ where

$$A_i^*(\theta_i) = \sum_{j=1}^{n} d_{ij}^* a(R(\hat{e}_{R,j} - \theta_i d_{ij}^*)). \tag{3.7}$$

Thus the problem has been reduced to finding n simple regressions. Furthermore these regressions are easily obtained. If we view the LHS of (3.7) as a function of θ_i, it can be shown that it is a decreasing step function of θ_i. The solution follows quickly using a simple linear search routine. A procedure which works quite well is the Illinois version of regula falsi similar to the algorithm discussed by McKean and Ryan (1977).

The R-residuals from the fit of the second form of the mean shift outlier model (A.3) are

$$\hat{\underline{e}}_R^* = \hat{\underline{e}}_R - \hat{\theta}_{R,i}\underline{d}_i^* \tag{3.8}$$

Define $\hat{\tau}(i)$ and $\hat{\tau}^*(i)$ as the estimates of τ and τ^* based on the residual vector $\hat{\underline{e}}_R^*$.

Note that if we replace the above rank criterion by the least squares criterion then we obtain the least squares estimate of θ_i by using a series of simple regressions to find a multiple regression; see Draper and Smith (1981, p.204).

3.3. RDFFIT.

Next we consider a statistic that measures the first order change in the R-fit of the ith case when the ith case is deleted. As in Appendix B, the first order terms in the change in the R-fit of the ith case when the ith case is deleted is

$$\begin{aligned} \text{RDFFIT}_i &= \hat{Y}_{R,i} - \hat{Y}_R(i) \\ &= \hat{\theta}_{R,i} h_i. \end{aligned} \tag{3.9}$$

Equation (3.9) can be developed as follows: For the ith case, consider the second formulation of the mean shift outlier model given by (A.3).

Appealing to the asymptotic orthogonality, $\hat{Y}_{R,i}$, the R-fit of Y_i in the original model (A.1), is the R-fit of the first term on the RHS of model (A.3) and $\hat{\theta}_{R,i}d_{ii}^{*}$ is the R-fit of the second term on the RHS of model (A.3). Hence the R-predicted value of Y_i in the mean shift outlier model is, to the first order, $\hat{Y}_{R,i} + \hat{\theta}_{R,i}d_{ii}^{*}$, which can be expressed as

$$\hat{Y}_{R,i} + \hat{\theta}_{R,i}d_{ii}^{*} = [\hat{Y}_{R,i} - \hat{\theta}_{R,i}(H\underline{d}_i)_i] + \hat{\theta}_{R,i}$$

$$= [\hat{Y}_{R,i} - \hat{\theta}_{R,i}h_i] + \hat{\theta}_{R,i}.$$

The term in brackets is, of course, the R-fit of the first term on the RHS of the first formulation of the mean shift outlier model, namely $(X\underline{b})_i$ of model (A.2). As noted in Appendix A, when least squares methods are used, the least squares fit of this term is $\hat{Y}_{LS}(i)$. Similar to least squares, the bracketed term

$$\hat{Y}_R(i) \doteq \hat{Y}_{R,i} - \hat{\theta}_{R,i}h_i. \tag{3.10}$$

Clearly, in order to be useful, RDFFIT_i needs to be assessed relative to some scale. The following R-diagnostics are formulations of RDFFITS_i based on appropriate scales.

3.4. RDCOOK and RDFFITS.

RDFFIT is a change in the fitted value; hence, a natural scale for assessing RDFFIT is a fitted value scale. It follows from Appendix B, see (B.5) and (B.6), that for the R-fit, assuming an asymmetric error distribution,

$$\text{Var}(\hat{Y}_{R,i}) \doteq \frac{1}{n}\tau^{*2} + h_{ic}\tau^{2}.$$

Hence, based on a fitted scale assessment, we standardize RDFFIT by

$(\mathrm{var}(\hat{Y}_{R,i}))^{1/2}$.

As noted in Appendix A, for least squares diagnostics there is some discussion on whether to use the original model or the mean shift outlier model for the estimation of scale. Cook and Weisberg (1982) advocate the original model. In this case the scale estimate is the same for all n cases. This allows casewise comparisons involving the diagnostic. Belsley, Kuh, and Welsch (1980), however, advocate scale estimation based on the mean shift outlier model. Note that both standardizations correct for the model and the underlying variation of the errors.

Let $\hat{\tau}^*$ and $\hat{\tau}$ denote the estimators of τ^* and τ discussed in Section 2. Our diagnostic in which RDFFIT_i is assessed relative to a fitted value scale with estimates of scale based on the original model is given by

$$(\mathrm{p\ RDCOOK}_i)^{1/2} = \frac{\mathrm{RDFFIT}_i}{(\frac{1}{n}\hat{\tau}^{*2}(i)+h_{ic}\hat{\tau}^2)^{1/2}}.$$

This is an R-analogue of $(\mathrm{p\ DCOOK}_i)^{1/2}$ statistic proposed by Cook and Weisberg (1982), see (A.9).

Let $\hat{\tau}^*(i)$ and $\hat{\tau}(i)$ denote the estimates of τ^* and τ for the mean shift outlier model as discussed above. Then our diagnostic in which RDFFIT_i is assessed relative to a fitted value scale with estimates of scale based on the mean shift outlier model is given by

$$\mathrm{RDFFITS}_i = \frac{\mathrm{RDFFIT}_i}{(\frac{1}{n}\hat{\tau}^{*2}(i)+h_{ic}\hat{\tau}^2(i))^{1/2}}. \qquad (3.11)$$

This is an R-analogue of the least squares diagnostic DFFITS_i proposed by Belsley et al. (1980); see (A.10) of Appendix A.

If the error distribution is assumed to be symmetric, the R-diagnostics are obtained by replacing $\mathrm{Var}(\hat{Y}_{R,i})$ with

$$\mathrm{Var}(\hat{Y}_{R,i}) = h_i\hat{\tau}^2,$$

see (B.6) of Appendix B. This eliminates the need to estimate τ^*.

There is disagreement on what cutoff values to use for flagging points of potential influence. As Belsley et al. (1980) discuss in some detail, DFFITS is inversely influenced by sample size. They advocate a size-adjusted cutoff value of $2\sqrt{p/n}$ for DFFITS, which would lead to a $2/\sqrt{n}$ cutoff value for $(DCOOK)^{1/2}$. Cook and Weisberg (1982, p.118) suggest a more conservative cutoff value of 1. In the examples in Section 4, we will use the more liberal value, realizing these diagnostics are only flagging potential influential points that require investigation. As with the two references cited above, we would never recommend indiscriminant deleting of observations solely because their diagnostic values exceed the cutoff point. Rather these are potential points of influence which should be investigated.

3.5. External t_R-statistic.

The above diagnostics RDCOOK and RDFITS, assess the first order change in RDFFIT relative to the R-fitted scale which is a τ-scale, (or a τ and τ^* scale under the assumption of an asymmetric error distribution). This change in fit, however, is proportional to $\hat{\theta}_{R,i}$. Hence assessing $\hat{\theta}_{R,i}$ on the τ-scale is consistent with the scale suggested by the approximate distribution of an R-estimate, see (2.6).

Note, that in the mean shift outlier model the leverage value of the ith case is 1. As Huber (1981) showed, a necessary and sufficient condition for the least squares estimates to be asymptotically normal is that the leverage values go to zero uniformly. Similarly, this is a sufficient condition for the asymptotic distribution theory of the R-estimates. Therefore the asymptotic theory for neither least squares nor R-estimates hold for the mean shift outlier model. Nevertheless, the external t-statistic, $t_{LS}(i)$, ($\hat{\theta}_{LS}$ relative to its standard error), see (A.8), has proved to be an effective diagnostic for least squares fit.

In analogy to the external t_{LS}-statistic, we propose the external t_R-statistic which is given by

$$t_R(i) = \frac{\hat{\theta}_{R,i}}{\hat{\tau}_{(i)}/\sqrt{1-h_i}}. \qquad (3.12)$$

Although this is the standardization suggested by the asymptotic distribution theory, in light of the above discussion, we do not propose it as a test for H_0: $\theta=0$ versus H_A: $\theta\neq 0$. Instead we propose it as an alternative to the least squares diagnostic $t_{LS}(i)$. We are still assessing the change RDFFIT on a τ-scale. We further feel that $\hat{\tau}$ is a more robust estimate of τ than $\hat{\sigma}$ is of σ and we have found in practice that it appears to be better at flagging potential points of influence than $t_{LS}(i)$.

3.6. RDFBETAS.

When the diagnostics RDFFITS or RDCOOK are large for, say, the ith case, then we usually want to investigate the impact this case has on the individual regression coefficients. Thus, we want to consider the statistic we shall define as

$$\text{RDFBETA}_i = \hat{\underline{b}}_R - \hat{\underline{b}}_R(i)$$

where $\hat{\underline{b}}_{R,i}$ is the R-estimate of $\underline{b}$ in the mean shift model (A.2).

In order to obtain this statistic, first note that if $\hat{\underline{Y}}_R$ is the R-prediction of $\underline{Y}$ in the original model, then the R-estimate of $\underline{b}$ is the solution $\hat{\underline{b}}_R$ to the equation

$$X\hat{\underline{b}}_R = \hat{\underline{Y}}_R$$

that is,

$$\hat{\underline{b}}_R = (X'X)^{-1}X'\hat{\underline{Y}}_R. \qquad (3.13)$$

In fact, most modern software obtains $\hat{\underline{b}}_R$ by first finding $\hat{\underline{Y}}_R$ employing a convenient basis matrix of X; see, for example, Hettmansperger and McKean (1983, Section 4).

Let $\chi = [X|\underline{d}_i]$ denote the design matrix for the mean shift outlier model. Let $\hat{\underline{Y}}_{R,i}$ denote the R-fitted value of this model. Then according to (3.13) $\hat{\underline{b}}_R(i)$ is the first p coordinates of the vector,

$$(\chi'\chi)^{-1}\chi'\hat{\underline{Y}}_{R,i}.$$

From Section 3.2, (3.10), $\hat{\underline{Y}}_{R,i} \doteq \hat{\underline{Y}}_R + \hat{\theta}_{R,i}\underline{d}_i^*$. Then using the result for the inverse of a partitioned matrix (see p.27 of Searle (1971)) and the fact that $\underline{d}_i^* = (I-H)\underline{d}_i$, we obtain after some algebra that

$$\hat{\underline{b}}_R(i) \doteq \hat{\underline{b}}_R - (X'X)^{-1}\chi_i'\hat{\theta}_{R,i}$$

where χ_i' is the ith row of χ. Hence

$$\text{RDFBETA}_i = (X'X)^{-1}\chi_i'\hat{\theta}_{R,i}.$$

To be useful RDFBETA_i needs to be measured relative to a scale. Since it is proportional to a difference in fitted values we shall choose a τ-scale. As in Section (3.4) if τ is estimated by using the mean shift outlier model. Then the diagnostic, defined for the jth component of RDFBETA_i, is

$$\text{RDFBETAS}_{i,j} = \text{RDFBETA}_i/(\hat{\tau}(i)\sqrt{(X_C'X_C)^{-1}_{jj}})$$

Belsley et al. (1980) advocate a size adjusted cutoff value of $2/\sqrt{n}$ for the corresponding least squares diagnostic.

These diagnostics are straightforward to compute. Consider the diagonal matrix $G_R = \text{diag}\{\hat{\theta}_{R,1},\ldots,\hat{\theta}_{R,n}\}$. Define $(p-1)\times n$ matrix

$$\text{RDFBETA} = [\hat{\underline{b}}_R - \hat{B}_R(1) \ \ldots\ \hat{\underline{b}}_R - \hat{\underline{b}}_R(n)].$$

It then follows

$$\text{RDFBETA} = (X_C'X_C)^{-1}X_C'G_R.$$

Note that each of the n-columns of RDFBETA is simply a least squares fit of a column of G. They can be obtained quickly using the QR-subroutines in

LINPACK; see Dongara et al. (1979).

4. Examples

The following two examples illustrate the power of the R-diagnostics in detecting influential points in linear models. The R-estimates were computed by the algorithm discussed in Hettmansperger and McKean (1983). To compute the diagnostics we used the LINPACK subroutines SQRDC and SQRSL for the numerical linear algebra parts, such as leverage values, projections, etc. The R-estimate of θ_i was computed as discussed in Section 3. The parameter τ was estimated as discussed in Koul et al. (1987) using the value of $\alpha = .80$.

Example 1. The data for this example can be found in Morrison (1983, p.64). The response is the level of free fatty acid of prepubescent boys while the independent variables are age, weight, and skin fold thickness. The sample size is 41. Figure 1 depicts the residual plot based on the least squares fit. From this plot there appears to be several outliers. Certainly the points 12, 22, 26 and 9 are outlying and perhaps the points 8, 10 and 38. In fact, the first four of these points probably spoiled the least squares fit, obscuring the points 8, 10 and 38. This seems apparent from the residual plot based on the Wilcoxon fit, Figure 2, where all seven points stand out.

Table 1 gives the values of the internal t, external t, DFFITS and $(DCOOK)^{1/2}$, diagnostic statistics for both the least squares and Wilcoxon fit. Using a cutoff value of 2 for the external t statistics and the suggested cutoff values of .62 for DFFITS and .31 for $(DCOOK)^{1/2}$, the least squares diagnostics flag only points 12 and 22 while the R-diagnostics flag all seven points. Both R-diagnostics are necessary; for instance RDFFIT and $(RDCOOK)^{1/2}$ flag point 8 while the external t_R is at 1.84. Conversely the external t_R flags point 26 while the other two do not.

Table 2 displays the RDFBETAS. Using the suggested cutoff value of .31, these statistics indicate an influential effect on at least one β for five of the above points and on the two exceptions, points 26 and 22, the outcome is borderline. Note that RDFFIT is large for β_1 at point 11.

Although this point was not flagged above, it is a point of high leverage; i.e. $h_{11} > 2p/n$.

Note that in both residual plots, the low values of the residuals are bunched together while the higher values are more dispersed; i.e., the distribution of the residuals appears to be positively skewed. For a final fit, then, we proceeded to use the bent R-score function given by

$$\varphi(u) = \begin{cases} \frac{3}{2} & \frac{3}{4} \le u \le 1 \\ \frac{14}{3}u-2 & 0 \le u \le \frac{3}{4} \end{cases}$$

which is suited for positively skewed error distributions with heavy right tails; see McKean and Sievers (1988) for a discussion of these scores. In its residual plot, Figure 3, the outliers stand out more than in the previous fits and it does appear to be more scattered indicating a better fit. The regression estimates for all three fits appear in Table 3. They do differ, especially the estimates of β_3. Table 4 displays the diagnostics for the bent score fit. Note that the above seven points are flagged as well as point 11.

Example 2. The dataset of this example is the stack-loss data presented in Daniel and Wood (1971, p.60). It has been discussed in several articles on robust methods, for instance, Andrews (1974) and Hettmansperger and McKean (1977). In the latter article, robust residuals plots are presented for fits using various R-scores. It appears from these plots that observations 1, 3, 4, and 21 are outlying points.

In Table 5 we present the diagnostic measures for both an R and a least squares fit, (the R-fit used Wilcoxon scores). The R-diagnostics clearly indicate that these points need further investigation. RDFFIT exceeds $2(p/n)^{1/2} = .87$ on all 4 of these points, the external t exceeds 2.0 on all but the first point (but even here it is at 1.91), and $(\text{RDCOOK})^{1/2}$ exceeds $2/\sqrt{n} = .44$ on points 1 and 21. From the RDFBETA values, points 1 and 3 had an impact on β_1 while the remaining two points had an impact on both β_1 and β_2. None of the R-diagnostics for the remaining 17 points exceeded the cutoff

values.

In contrast, for least squares, only observation 21 was flagged by DFFITS while observations 4 and 21 were flagged by the least squares external t statistic. The remaining two were not flagged.

5. Conclusions

Diagnostics are an extremely important part of many data analyses. Least squares diagnostics have been effective in detecting and identifying aberrant cases. These methods fit most naturally with least squares based inference. Currently, there are several approaches to robust inference in the linear model. The present paper suggests natural diagnostic quantities to be used in conjunction with robust rank-based inference. The robust diagnostics appear to have some advantages. In the examples, they were able to flag cases of potential trouble that were passed over by least squares diagnostics.

Appendix A.

In this appendix we derive the least squares diagnostic tools (internal and external t, DFFITS, DCOOK, and DFBETAS) from a common source (the mean shift outlier model). We also establish some of the results we need in the derivation of the R-diagnostics.

Consider the linear model,

$$\underline{Y} = X\underline{b} + e \qquad (A.1)$$

which is defined in Section 2. The mean shift outlier model for the ith data point is defined by

$$\underline{Y} = X\underline{b} + \underline{d}_i\theta_i + \underline{e} \qquad (A.2)$$

where $\underline{d}_i$ is a $n\times 1$ vector of zeroes except its ith component is 1. The parameters θ_i, $i = 1,\ldots,n$, play a key role in the diagnostics.

There are several ways of writing model (A.2). Following Cook and Weisberg (1982) and letting $\underline{d}_i^* = (I-H)\underline{d}_i$, where H is $X(X'X)^{-1}X'$, the model can

be written as

$$\underline{Y} = X\underline{b}^* + \underline{d}_i^*\theta_i + \underline{e}. \tag{A.3}$$

Since X and $\underline{d}_i^*$ are orthogonal, the least squares estimate of θ_i is

$$\hat{\theta}_i = \frac{\underline{d}_i^{*\prime}\underline{Y}}{\underline{d}_i^{*\prime}\underline{d}_i^*} = \frac{\hat{e}_{LS,i}}{1-h_{ii}} . \tag{A.4}$$

The second equality holds since $\underline{d}_i^{*\prime}\underline{d}_i^* = 1-h_{ii}$ and $d_i^{*\prime}\underline{Y} = d_i'(I-H)\underline{Y}$.

Next we want to connect $\hat{\theta}_i$ with the statistic $DFFIT_i$ which is the difference in the fitted value of Y_i at the model (A.1) and when the ith point is deleted. Let $\hat{Y}_{LS,i}$ be the fitted value of Y_i at model (A.1) and let $\hat{Y}_{LS}(i)$ be the fitted value of Y_i when the ith point is deleted. Then

$$DFFIT_i = \hat{Y}_{LS,i} = \hat{Y}_{LS}(i).$$

In order to obtain $\hat{Y}_{LS}(i)$, we need not delete the point and refit since it follows from Cook and Weisberg (1982, p.33) that

$$\hat{Y}_{LS}(i) = Y_i - \hat{\theta}_i;$$

hence,

$$\begin{aligned} DFFIT_i &= \hat{Y}_{LS,i} - (Y_i - \hat{\theta}_i) \\ &= -(1-h_{ii})\hat{\theta}_i + \hat{\theta}_i \\ &= h_{ii}\hat{\theta}_i \end{aligned} \tag{A.5}$$

where the middle equality follows from (A.4).

The least squares diagnostics follow from different standardizations of $DFFIT_i$. For the t-statistics note from (A.4) that,

$$Var(\hat{\theta}_i) = \sigma^2/(1-h_{ii}).$$

If we standardize $DFFIT_i$ by using the estimate $\hat{\sigma}^2$ of σ^2 based on model (A.1) and use (A.4) we then get the internal t statistic given by

$$\frac{DFFIT_i}{h_{ii}(\hat{\sigma}/\sqrt{1-h_{ii}})} = \frac{\hat{\theta}_i}{\hat{\sigma}/\sqrt{1-h_{ii}}} = \frac{\hat{e}_{LS,i}}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{A.6}$$

$$= r_{LS,i}$$

This is called the internally studentized residual; see Cook and Weisberg (1982, p.18).

Next suppose we standardize $DFFIT_i$ by the estimate $s^2(i)$ of σ^2 based on the mean shift outlier model (A.2). As derived in Cook and Weisberg (1982, p.20)

$$s^2(i) = \frac{(n-p-1)\hat{\sigma}^2 - \hat{e}_i^2/(1-h_{ii})}{n-p-2} . \tag{A.7}$$

The corresponding standardization of $DFFIT_i$ is,

$$\frac{DFFIT_i}{h_{ii}(s(i)/\sqrt{1-h_{ii}})} = \frac{\hat{\theta}_i}{s(i)/\sqrt{1-h_{ii}}} \tag{A.8}$$

$$= \hat{t}_{LS}(i)$$

This is the externally studentized residual; see Cook and Weisberg (1982, p.20). This is also the t-statistic for testing H_0: $\theta_i=0$ versus H_A: $\theta\neq\theta_i$ in model A.2.

The above standardizations of $DFFIT_i$ are consequences of considering it in terms of $\hat{\theta}_i$. Suppose instead we standardize it in terms of fitted values. Note that

$$Var(\hat{Y}_i) = \sigma^2 h_{ii}.$$

If we standardize $DFFIT_i$ by using $\hat{\sigma}^2$ as our estimate of σ^2 we get

$$\frac{DFFIT_i}{\hat{\sigma}\sqrt{h_{ii}}} = \frac{\hat{\theta}\sqrt{h_{ii}}}{\hat{\sigma}}$$

$$= r_{LS,i}\sqrt{\frac{h_{ii}}{1-h_{ii}}} \tag{A.9}$$

$$= \sqrt{p\ \text{DCOOK}}.$$

These equalities follow from (A.4) and (A.6). See Cook and Weisberg (1982, p.117).

If on the other hand, we standardize it by using $s^2(i)$ as our estimate of σ^2 we get

$$\frac{\text{DFFIT}_i}{s(i)\sqrt{h_{ii}}} = \frac{\hat{\theta}_i\sqrt{h_{ii}}}{s(i)} \qquad (A.10)$$

$$= \text{DFFITS}_i.$$

This statistic was proposed by Belsley et al. (1980).

Before considering DFBETA, for Section 3 we need the following result. Under the mean shift outlier model (A.2), it follows immediately that the predicted value of the ith observation is Y_i. Hence under this model we have

$$Y_i = (X\hat{\underline{b}})_{LS,i} + \hat{\theta}_i$$

where $(X\hat{\underline{b}})_{LS,i}$ is the least squares estimate of the first term on the RHS of (A.2). Since $\hat{\theta}_i = Y_i - \hat{Y}_{LS}(i)$ we have that

$$\hat{Y}_{LS}(i) = (X\hat{\underline{b}})_{LS,i}.$$

Thus $\hat{Y}_{LS}(i)$ is also the least squares estimate of the first term on the RHS of model (A.2).

When DFFITS is large for, say the ith point, usually we want to investigate its impact on the individual coordinates. Consider the diagonal matrix $G = \text{diag}\{\hat{\theta}_1,\ldots,\hat{\theta}_n\}$. Let $\hat{\underline{b}}_{LS}(i)$ denote the estimate of $\underline{b}$ when the ith point is deleted. Then the $(p-1)\times n$ matrix of changes in the coordinates is given by

$$\text{DFBETA} = [\hat{\underline{b}}-\hat{\underline{b}}_{LS}(1)\ldots\hat{\underline{b}}-\hat{\underline{b}}_{LS}(n)].$$

It follows from Belsley et al. (1980, p.13) that

$$\text{DFBETA} = (X'X)^{-1}X'G.$$

These can be obtained quickly using the QR-subroutines in LINPACK.

As with $DFFIT_i$, we can standardize DFBETA in several ways. The one we shall note here is to use s(i) to estimate σ. This leads to

$$DFBETS_{LSi,j} = \frac{\hat{b}_j - \hat{b}_j(i)}{s(i)\sqrt{(X'X)^{-1}_{jj}}} = \frac{(DFBETA)_{j,i}}{s(i)\sqrt{(X'X)^{-1}_{jj}}} .$$

which a diagnostic proposed by Belsley et al. (1980, p.13).

Appendix B

In this appendix, we develop the approximations up to terms of order n^{-1} for the variance-covariance matrices $Cov(\hat{\underline{e}}_R)$ and $Cov(\hat{\underline{y}}_R)$. We will concentrate on the case of asymmetric error distributions and state the results in the symmetric case. We will use the notation $H_c = P_{X_c} = X_c(X_c'X_c)^{-1}X_c'$ and $J = P_1 = n^{-1}(\underline{1}\ \underline{1}')$ along with h_{ic} the ith diagonal element of H_c. Then the leverage of the ith case is $h_i = n^{-1} + h_{ic}$. The main results are

$$Cov(\hat{\underline{e}}_R) \doteq \sigma^2\{I - K_1 J - K_2 H_c\} \tag{B.1}$$

where $K_1 = (\tau^*/\sigma)^2(2\delta^*/\tau^* - 1)$ (B.2)

$$K_2 = (\tau/\sigma)^2(2\delta/\tau - 1)$$

$$\delta^* = E(e_i \operatorname{sgn} e_i)$$

$$\delta = E[e_i a(F(e_i))]$$

σ^2 is the error variance and τ^*, τ defined in Section 2.

Hence,

$$Var\ \hat{e}_{R,i} \doteq \sigma^2(1 - K_1 n^{-1} - K_2 h_{ic}). \tag{B.3}$$

In the case of a symmetric error distribution,

$$Var\ \hat{e}_{R,i} \doteq \sigma^2(1 - K_2 h_i). \tag{B.4}$$

Recall Cook and Weisberg (1982, p.11), that in the least squares case, Var $\hat{e}_{LS,i} = \sigma^2(1-h_i)$ so that K_1 and K_2 are correction factors due to using the rank score fitting algorithm.

Likewise

$$\text{Cov}(\hat{\underline{Y}}_R) \doteq \tau^{*2} J + \tau^2 H \tag{B.5}$$

$$\text{Var } \hat{Y}_{Ri} \doteq n^{-1}\tau^{*2} + h_{ic}\tau^2,$$

while in the symmetric case,

$$\text{Var } \hat{Y}_{Ri} \doteq h_i \tau^2. \tag{B.6}$$

Before giving a sketch of the derivations of these formulas, we discuss the estimation of the parameters appearing above. Natural estimates of δ^* and δ are

$$\hat{\delta}^* = \frac{1}{n-p} \sum_{i=1}^{n} |\hat{e}_{R,i}| \tag{B.7}$$

$$\hat{\delta} = \frac{1}{n-p} \sum_{i=1}^{n} \hat{e}_i a(R(\hat{e}_i))$$

$$= \frac{1}{n-p} D(\hat{\beta}_R), \tag{B.8}$$

where $D(\beta)$ is defined in (2.2). Estimates of τ^* and τ are referenced in Section 2.

We now outline an approximation for $\text{Cov}(\hat{\underline{e}}_R)$, the variance-covariance matrix of $\hat{\underline{e}}_R$, the vector of residuals. Using (2.4) and (2.8),

$$\hat{\underline{e}}_R \doteq Y - \underline{1}(\alpha+\tau^* n^{-1} \underline{1}\, \underline{a}^*) - X_c(\beta+\tau(X_c'X_c)X_c'\underline{a})$$

where $\underline{a} = \underline{a}(R(\hat{\underline{e}}))$ and $\underline{a}^{*\prime} = (\text{sgn } \hat{e}_1, \ldots, \text{sgn } \hat{e}_n)$.

Then

$$\hat{\underline{e}}_R \doteq \underline{e} - \tau^* J\underline{a}^* - \tau H_c \underline{a}.$$

Now $E\underline{a} \doteq \underline{0} \doteq E\underline{a}^*$ and hence

$$\text{Cov}(\hat{\underline{e}}_R) \doteq E\{\underline{e}\,\underline{e}' - 2\tau^*\underline{e}\,\underline{a}^{*\prime}J - 2\tau\underline{e}\,\underline{a}'H_C$$
$$+ \tau^{*2}J\underline{a}^*\underline{a}^{*\prime}J + 2\tau^*\tau J\underline{a}^*\underline{a}'H_C$$
$$+ \tau^2 H_C\underline{a}\,\underline{a}'H_C\}.$$

Note that $E\underline{a}\,\underline{a}' \doteq I \doteq Ea^*a^{*\prime}$. Further $E\underline{e}\,\underline{a}^{*\prime} \doteq \delta^* I$, $E\underline{e}\,\underline{a}' \doteq \delta I$, and $E\underline{a}^*\underline{a}' \doteq cI$ for a constant c. Now using $J'J = J$, $H_C'H_C = H_C$, $J'H_C = 0$ we have

$$\text{Cov}(\hat{\underline{e}}_R) \doteq \sigma^2 I - \tau^{*2}(\frac{2\delta}{\tau^*} - 1)J - \tau^2(\frac{2\delta}{\tau} - 1)H_C.$$

Then (B.1), (B.2) and (B.3) follow immediately. The formula (B.5) follows in a similar fashion from $\hat{\underline{Y}} \doteq \underline{1}\alpha + X_C\underline{\beta} + \tau^* J\underline{a}^* + \tau H_C\underline{a}$. Similarly for formulas (B.4) and (B.6).

References

Andrews, D. F. (1974). A robust method for multiple linear regression. Technometrics, 16, 523-531.

Aubuchon, J. C. and Hettmansperger, T. P. (1988). Rank-based inference for linear models: Asymmetric errors. Stat. and Prob. Letters. To appear.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics. John Wiley, New York.

Chatterjee, S. and Hadi, A. S. (1986). Influential observations (with discussion). Statistical Science, 1, 379-416.

Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, New York.

Cook, R. D. (1986). Assessment of local influence (with discussion). J. R. Statist. Soc. B, 48, 133-169.

Daniel, C. and Wood, F. (1971). Fitting Equations to Data. John Wiley, New York.

Dongara, J. J., Bunch, J. R., Moler, C. B. and Stewart, G. N. (1979). LINPACK Users Guide. SIAM, Philadelphia.

Draper, N. R. and Smith, H. (1981). Applied Regression Analysis, 2nd ed. John Wiley, New York.

Heiler, S. and Willers, R. (1979). Asymptotic normality of R-estimates in the linear model. Forschungsbericht NR.79/6, University of Dortmund, Dortmund.

Hettmansperger, T. P. (1984). Statistical Inference Based on Ranks. John Wiley, New York.

Hettmansperger, T. P. (1987). Why not try a robust regression? Austral. J. Statist., 29, 1-18.

Hettmansperger, T. P. and McKean, J. W. (1977). A robust alternative based on ranks to least squares in analyzing general linear models. Technometrics, 19, 275-284.

Hettmansperger, T. P. and McKean, J. W. (1983). A geometric interpretation of inferences based on ranks in the linear model. J. Am. Statist. Assoc., 78, 885-893.

Hocking, R. R. (1983). Developments in linear regression methodology 1959-1982. Technometrics, 25, 219-230, with discussion.

Huber, P. J. (1981). Robust Statistics. John Wiley, New York.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. Ann. Math. Statist., 43, 1449-1458.

Koul, H. L., Sievers, G. L. and McKean, J. W. (1987). An estimator of the scale parameter for the rank analysis of linear models under general score functions. Scand. J. Statist., 14, 131-141.

McKean, J. W. (1975). Tests of Hypotheses Based on Ranks in the General Linear Model, Unpublished Ph.D. Dissertation. Penn State University, University Park, PA.

McKean, J. W. and Hettmansperger, T. P. (1976). Tests of hypotheses in the general linear model based on ranks. Comm. in Statist., A5, 693-709.

McKean, J. W. and Hettmansperger, T. P. (1978). A robust analysis of the general linear model based on one step R-estimates. Biometrika, 65, 571-579.

McKean, J. W. and Ryan, T. A. Jr. (1977). An algorithm for obtaining confidence intervals and point estimates based on ranks in the two sample location problem. Trans. Math. Software, 3, 183-185.

McKean, J. W. and Schrader, R. M. (1984). A comparison of methods for studentizing the sample median. Comm. Statist.-Simulation and Computation, 13(6), 751-773.

McKean, J. W. and Sievers, G. L. (1988). Rank scores suitable for analyses of linear models under asymmetric error distributions. Unpublished manuscript.

Morrison, D. F. (1983). Applied Linear Statistical Models. Prentice-Hall, Englewood Cliffs, NJ.

Osborne, M. R. (1985). Finite Algorithms in Optimization and Data Analysis. John Wiley, New York.

Puri, M. L. and Sen, P. K. (1985). Nonparametric Methods in General Linear Models. John Wiley, New York.

Searle, S. R. (1971). Linear Models. John Wiley, New York.

Shoemaker, L. H. and Hettmansperger, T. P. (1982). Robust estimates and tests for the one- and two-sample scale models. Biometrika, 69, 47-53.

Table 1. Diagnostics for R (Wilcoxon Scores)
and Least Squares fit of Example 1.

Case	Int(R)	Ext(R)	RDFFITS	RDCOOK	Int(LS)	Ext(LS)	DFFITS	DCOOK
1	0.72	0.80	0.20	0.10	0.49	0.49	0.12	0.06
2	-0.79	-1.00	-0.34	-0.17	-1.19	-1.20	-0.40	-0.20
3	-0.16	-0.19	-0.05	-0.03	-0.53	-0.53	-0.15	-0.07
4	-0.64	-0.84	-0.19	-0.10	-0.96	-0.95	-0.22	-0.11
5	0.69	0.76	0.28	0.14	0.42	0.42	0.15	0.08
6	0.12	0.13	0.03	0.02	-0.21	-0.21	-0.05	-0.02
7	-0.72	-0.92	-0.32	-0.17	-1.07	-1.07	-0.37	-0.18
8	1.51	1.84	0.67	0.32	1.16	1.17	0.43	0.21
9	2.07	2.69	0.56	0.26	1.74	1.79	0.38	0.18
10	1.54	1.94	0.62	0.29	1.12	1.13	0.36	0.18
11	0.93	1.05	0.53	0.25	0.56	0.56	0.30	0.15
12	3.18	4.13	1.03	0.47	2.84	3.17	0.79	0.35
13	-0.56	-0.65	-0.27	-0.14	-0.77	-0.77	-0.35	-0.18
14	0.00	-0.00	-0.00	-0.00	-0.34	-0.34	-0.09	-0.04
15	0.21	0.28	0.08	0.04	-0.12	-0.12	-0.04	-0.02
16	0.66	0.78	0.18	0.09	0.31	0.31	0.07	0.04
17	-0.03	-0.11	-0.02	-0.01	-0.34	-0.34	-0.07	-0.03
18	-0.72	-0.91	-0.26	-0.14	-1.12	-1.13	-0.32	-0.16
19	-0.43	-0.56	-0.24	-0.13	-0.83	-0.82	-0.36	-0.18
20	-0.66	-0.86	-0.30	-0.16	-1.00	-1.00	-0.35	-0.17
21	-0.60	-0.80	-0.13	-0.07	-0.90	-0.89	-0.15	-0.07
22	2.43	2.80	0.61	0.30	2.26	2.40	0.53	0.25
23	0.16	0.19	0.03	0.02	-0.11	-0.10	-0.02	-0.01
24	-0.72	-0.90	-0.22	-0.12	-1.09	-1.09	-0.27	-0.13
25	-0.32	-0.39	-0.14	-0.07	-0.49	-0.49	-0.18	-0.09
26	1.73	2.06	0.43	0.21	1.51	1.53	0.32	0.16
27	-0.76	-0.96	-0.19	-0.10	-1.09	-1.10	-0.21	-0.11
28	0.47	0.58	0.14	0.07	0.24	0.24	0.06	0.03
29	1.00	1.27	0.39	0.19	0.76	0.75	0.23	0.12
30	-0.51	-0.59	-0.18	-0.09	-0.71	-0.70	-0.22	-0.11
31	-0.78	-0.97	-0.20	-0.10	-1.05	-1.05	-0.22	-0.11
32	-0.19	-0.24	-0.07	-0.04	-0.36	-0.36	-0.11	-0.06
33	0.84	1.02	0.24	0.12	0.57	0.57	0.14	0.07
34	-0.58	-0.68	-0.15	-0.08	-0.81	-0.81	-0.18	-0.09
35	-0.69	-0.86	-0.23	-0.12	-0.98	-0.97	-0.26	-0.13
36	-0.02	-0.13	-0.13	-0.07	0.18	0.18	0.19	0.09
37	0.67	0.76	0.20	0.10	0.49	0.49	0.13	0.06
38	1.64	1.95	0.82	0.40	1.27	1.29	0.54	0.27
39	0.89	1.01	0.41	0.21	0.73	0.73	0.30	0.15
40	-0.30	-0.41	-0.21	-0.11	-0.51	-0.50	-0.29	-0.15
41	0.07	0.10	0.04	0.02	-0.09	-0.08	-0.03	-0.02

Table 2. RDFBETAS (Wilcoxon Scores) for Example 1.

Case	Incep.	Age	Weight	Skinfold
1	0.08	-0.08	-0.02	0.11
2	-0.09	0.20	-0.25	0.25
3	-0.04	0.02	0.02	-0.00
4	-0.13	0.07	0.06	-0.06
5	0.16	-0.11	-0.09	0.17
6	0.02	-0.02	0.00	0.00
7	-0.16	0.27	-0.20	0.09
8	0.43	-0.00	-0.49	0.21
9	0.34	-0.14	-0.08	-0.17
10	0.47	-0.17	-0.19	-0.15
11	0.14	-0.44	0.49	-0.27
12	0.83	-0.48	-0.15	-0.18
13	-0.09	-0.01	0.20	-0.26
14	-0.00	0.00	0.00	-0.00
15	0.06	-0.02	-0.05	0.03
16	0.06	-0.07	0.09	-0.13
17	-0.01	-0.00	0.01	-0.00
18	-0.09	0.09	-0.11	0.20
19	-0.02	0.13	-0.21	0.19
20	-0.10	-0.11	0.26	-0.09
21	-0.04	0.00	0.02	0.00
22	-0.16	0.04	0.05	0.29
23	0.01	-0.01	0.01	0.01
24	-0.04	0.07	-0.12	0.17
25	0.01	0.06	-0.07	-0.05
26	-0.06	-0.11	0.25	-0.06
27	-0.02	-0.02	0.00	0.08
28	-0.08	0.05	0.05	-0.04
29	-0.24	0.25	0.02	-0.14
30	0.04	-0.11	0.13	-0.12
31	0.08	-0.12	0.02	0.03
32	0.04	-0.01	-0.04	0.00
33	-0.06	0.16	-0.11	0.00
34	0.07	-0.10	0.05	-0.02
35	0.04	-0.16	0.13	-0.01
36	0.06	-0.01	-0.01	-0.10
37	-0.14	0.09	0.05	-0.01
38	-0.19	0.32	0.08	-0.57
39	-0.34	0.28	0.07	-0.07
40	0.12	-0.22	0.14	-0.05
41	-0.03	0.02	0.01	-0.01

Table 3 Fits for Example 1,
(standard error in parentheses).

Fit	Incep.	Age	Weight	SkinFold	Scale $\hat{\sigma}$ or $\hat{\tau}$
Least Squares	1.70 (.327)	-.0021 (.003)	-.0152 (.005)	.2045 (.166)	.215
R-Wilcoxon	1.49 (.273)	-.0011 (.003)	-.0154 (.004)	.2739 (.137)	.178
R-Bent Score	1.43 (.247)	-.0009 (.002)	-.0152 (.004)	.3079 (.124)	.159

Table 4. Diagnostics for R (Bent Scores) for Example 1.

Case	Int(R)	Ext(R)	RDFFITS	RDCOOK
1	0.67	1.07	0.27	0.14
2	-0.77	-0.97	-0.33	-0.17
3	-0.14	-0.13	-0.04	-0.02
4	-0.64	-0.77	-0.17	-0.09
5	0.64	0.97	0.35	0.18
6	0.11	0.32	0.07	0.04
7	-0.72	-0.94	-0.32	-0.16
8	1.53	2.16	0.79	0.38
9	2.09	3.14	0.66	0.30
10	1.58	2.57	0.82	0.35
11	0.89	1.13	0.61	0.24
12	3.20	5.20	1.29	0.53
13	-0.61	-0.71	-0.33	-0.17
14	0.01	0.18	0.05	0.02
15	0.22	0.46	0.13	0.07
16	0.67	1.07	0.25	0.13
17	-0.03	0.13	0.03	0.01
18	-0.68	-0.88	-0.25	-0.13
19	-0.41	-0.52	-0.23	-0.12
20	-0.63	-0.80	-0.28	-0.15
21	-0.61	-0.73	-0.12	-0.06
22	2.35	3.30	0.72	0.34
23	0.13	0.34	0.06	0.03
24	-0.71	-0.89	-0.22	-0.12
25	-0.41	-0.52	-0.19	-0.09
26	1.69	2.43	0.51	0.24
27	-0.75	-0.92	-0.18	-0.09
28	0.43	0.80	0.19	0.09
29	0.97	1.48	0.46	0.21
30	-0.57	-0.64	-0.20	-0.11
31	-0.80	-0.95	-0.20	-0.11
32	-0.27	-0.34	-0.10	-0.05
33	0.83	1.35	0.32	0.15
34	-0.61	-0.71	-0.16	-0.09
35	-0.69	-0.85	-0.22	-0.12
36	-0.35	-0.39	-0.39	-0.19
37	0.61	1.09	0.28	0.13
38	1.67	2.46	1.04	0.46
39	0.81	1.39	0.57	0.27
40	-0.32	-0.36	-0.19	-0.09
41	0.00	0.21	0.08	0.04

Table 5. Diagnostics for R (Wilcoxon Scores)
and Least Squares fit of Example .2.

Case	Int(R)	Ext(R)	RDFFITS	RDCOOK	Int(LS)	Ext(LS)	DFFITS	DCOOK
1	1.42	1.93	1.28	0.45	1.19	1.21	0.79	0.39
2	-0.50	-0.69	-0.48	-0.23	-0.72	-0.71	-0.48	-0.24
3	1.61	2.13	1.02	0.37	1.55	1.62	0.74	0.36
4	2.22	2.79	1.15	0.40	1.88	2.05	0.79	0.36
5	-0.45	-0.41	-0.12	-0.05	-0.54	-0.53	-0.12	-0.06
6	-0.75	-0.79	-0.24	-0.11	-0.97	-0.96	-0.28	-0.14
7	-0.55	-0.61	-0.32	-0.15	-0.83	-0.83	-0.44	-0.22
8	-0.20	-0.19	-0.10	-0.05	-0.48	-0.47	-0.25	-0.13
9	-0.71	-0.77	-0.31	-0.14	-1.05	-1.05	-0.42	-0.21
10	0.21	0.18	0.09	0.05	0.44	0.43	0.21	0.11
11	0.55	0.55	0.25	0.12	0.88	0.88	0.38	0.19
12	0.51	0.70	0.38	0.17	0.97	0.97	0.51	0.26
13	-0.72	-0.87	-0.38	-0.16	-0.48	-0.47	-0.20	-0.10
14	-0.28	-0.32	-0.16	-0.07	-0.02	-0.02	-0.01	-0.00
15	0.71	0.84	0.41	0.19	0.81	0.80	0.39	0.20
16	0.24	0.41	0.15	0.08	0.30	0.29	0.11	0.06
17	-0.33	-0.31	-0.26	-0.12	-0.61	-0.60	-0.50	-0.26
18	0.00	0.00	0.00	0.00	-0.15	-0.15	-0.07	-0.03
19	0.07	0.05	0.03	0.01	-0.20	-0.20	-0.09	-0.05
20	0.52	0.47	0.16	0.07	0.45	0.44	0.13	0.07
21	-3.15	-2.94	-1.92	-0.83	-2.64	-3.33	-2.10	-0.83

Figure 1. Residual plot for LS fit of Example 1

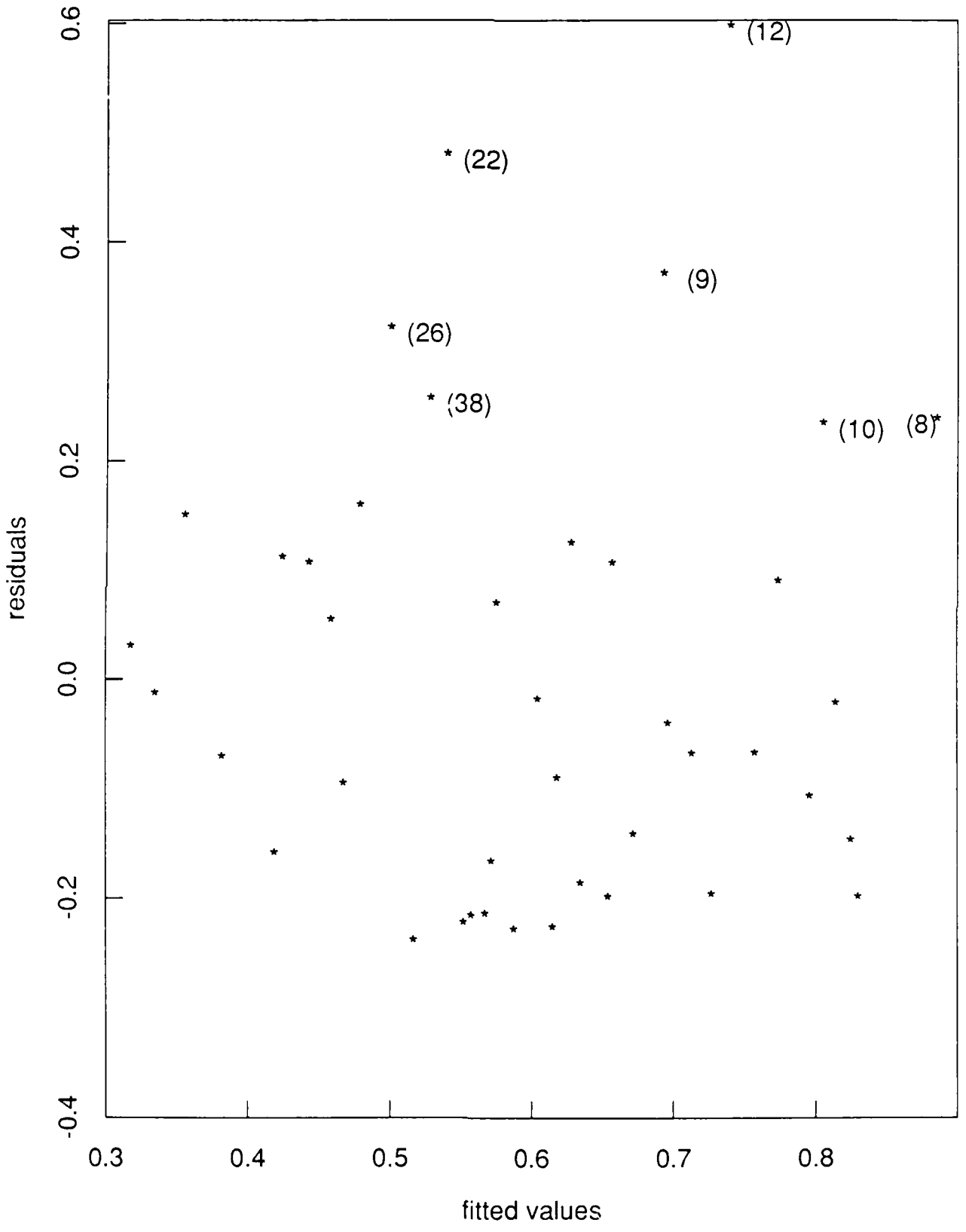


Figure 2. Residual plot for Wilcoxon fit of Example 1

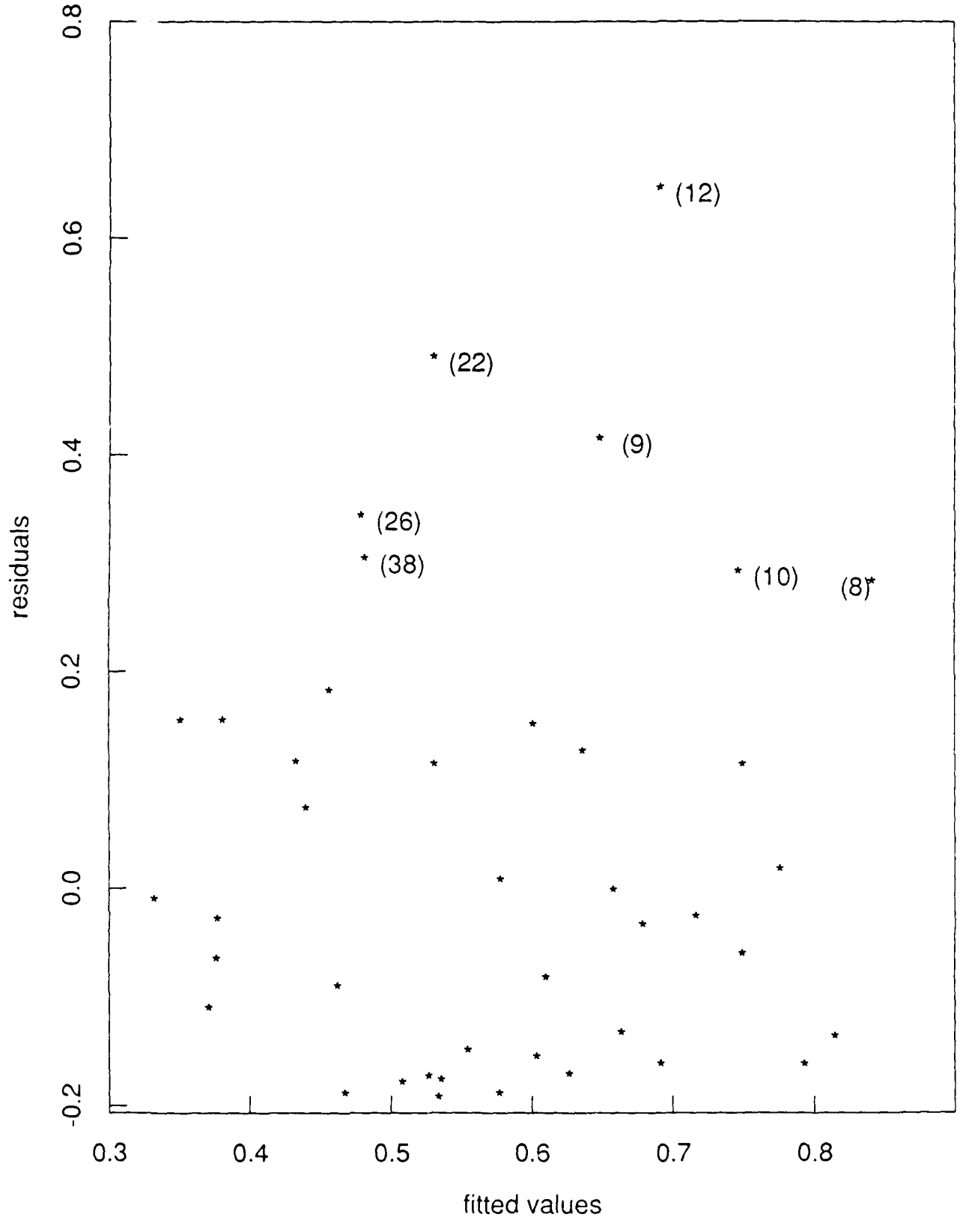
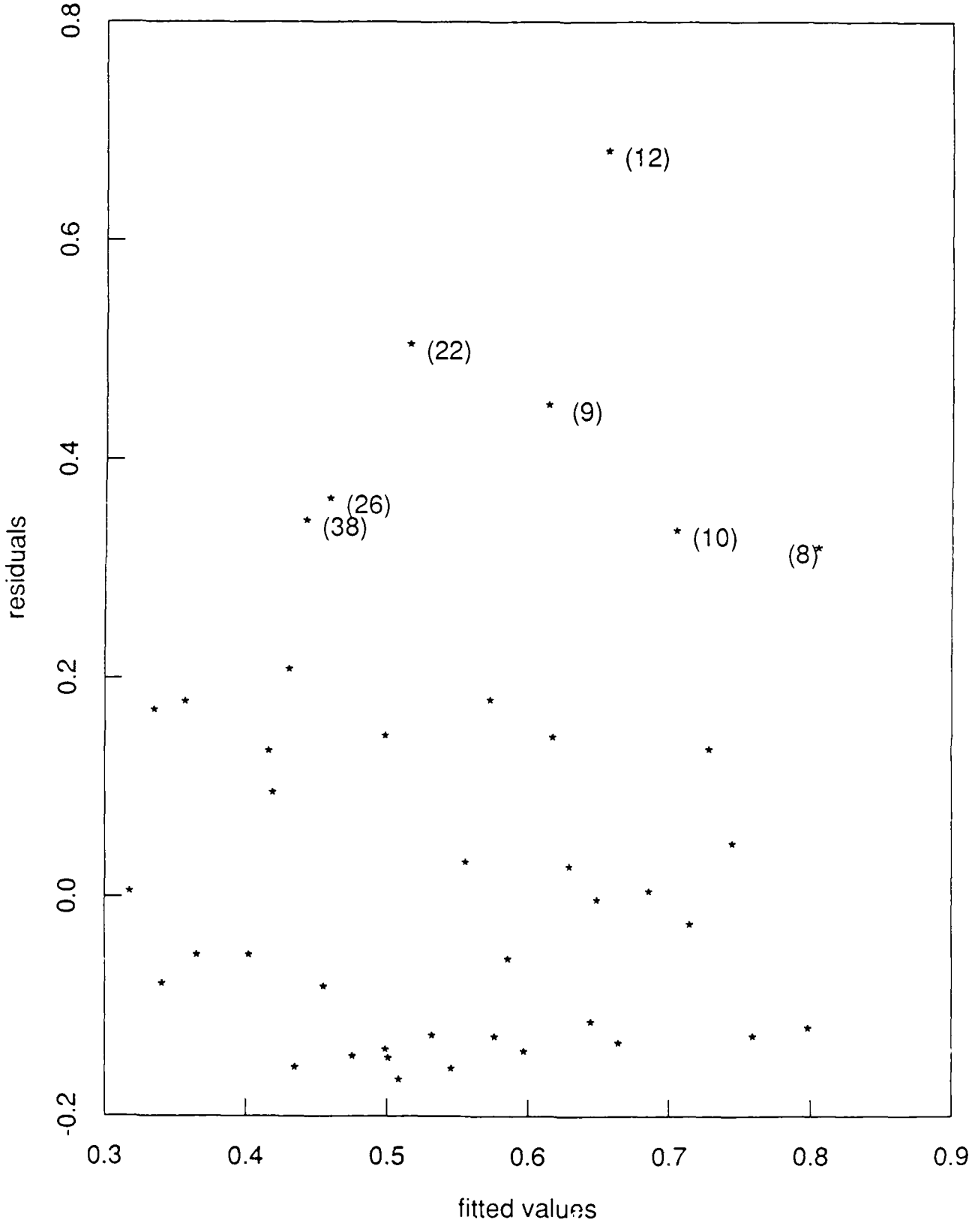


Figure 3. Residual plot for Bent Score fit of Example 1



Unclassified

ADA199834

REPORT DOCUMENTATION PAGE

1. REPORT NUMBER: 76

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): Robust Diagnostics for Rank-Based Inference

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Joseph W. McKean, Western Michigan University
Simon J. Sheather, Univ. of New South Wales
Thomas P. Hettmansperger, Pennsylvania State Univ.

8. CONTRACT OR GRANT NUMBER(s): N00014-80-C0741

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Department of Statistics
The Pennsylvania State University
University Park, PA 16802

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR042-446

11. CONTROLLING OFFICE NAME AND ADDRESS:
Office of Naval Research
Statistical and Probability Program Code 436
Arlington, VA 22217

12. REPORT DATE: September 1988

13. NUMBER OF PAGES: 28

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Linear Models, Robustness, Regression.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Diagnostics based on robust R-estimates of regression coefficients are developed. These methods are not as sensitive to influential points as least squares diagnostics. In data sets with several influential points, diagnostics based on a robust fit have a greater chance of detecting interesting cases for further inspection. Robust analogues of the internal and external t statistics, DFFITS, DCOOK, and DFBETAS are developed and illustrated on two data sets.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)